Matching Supply and Demand in Production-Inventory Systems: Asymptotics and Optimization

Yingdong Lu, Mark S. Squillante

Mathematical Sciences Department, IBM Research, Yorktown Heights, NY 10598, USA. {yingdong,mss}@us.ibm.com

David D. Yao

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA.
yao@ieor.columbia.edu

We consider a general class of high-volume, fast-moving production-inventory systems based on both lost-sales and backorder inventory models. Such systems require a fundamental understanding of the asymptotic behavior of key performance measures under various supply strategies, as well as the pre-planning of these strategies. Our analysis relies on a thorough study of the asymptotic behavior of a random walk with power drift, which is of independent interest. In addition to providing key insights, our analysis leads to approximations of the corresponding optimization problem that yield simple solutions which are close to optimal. We also establish an equivalence between the lost-sales and backorder models when both have the same penalty cost that becomes large.

1. Introduction

The production-inventory problem we study is that of matching supply and demand over a planning horizon of N periods, with the provision that a set of supply decisions must be made before the first period. For example, the production-inventory planning horizon might represent a "selling season" and the decisions to match supply with demand are made before the season starts. This is not merely a quantity decision — how much to stock up so as to fulfill demand over the entire season; rather, it is more like a capacity planning decision — how to set the supply or production rates as a function of time over the horizon, vis-à-vis the demand forecast. We consider this production-inventory matching problem within the context of both lost-sales and backorder inventory models.

More precisely, consider the following general supply and demand matching problem. Suppose $D_1 + \cdots + D_n$ denotes the cumulative demand up to period $n = 1, 2, \ldots, N$; and assume $\mathsf{Var}(D_n) = \sigma^2$ for all n. Further suppose the set of supply decisions before the first period takes the form $\mathsf{E}[D_1] + \cdots + \mathsf{E}[D_n] + \kappa \sigma n^\alpha$, where α and κ are decision variables such that $\alpha \in [0, 1]$ while κ can be either positive or negative. This assumed form allows a fair amount of flexibility: one can

choose to over-supply (positive κ), thus creating safety stock, if the lost-sales penalty is high; or to under-supply (negative κ), if the production cost or holding cost is high. We are especially interested in asymptotic results. When n is large, demand will approach a normal distribution, with a standard deviation equal to $\sigma\sqrt{n}$. Shall we match supply with demand by setting $\alpha = \frac{1}{2}$? What is the corresponding lost-sales quantity? What is the impact on lost sales if α takes on other values? What is the right value for κ; in particular, should it be positive or negative?

We note that this general supply and demand matching problem arises in a wide variety of production-inventory systems. As such, there is another way to interpret the asymptotics associated with large values of N in our model. Instead of viewing this as a prolonged planning horizon, we can view the latter as having a fixed length, but with the demand being scaled up. In other words, the original total demand for the entire season, D, now becomes $D_1 + \cdots + D_N$, with the D_n being independent copies of D. The capacity problem continues to consist of making the set of supply decisions for the entire horizon in advance, with the corresponding deliveries received every period to satisfy the demand. Indeed, this alternative view is closer to the class of production-inventory applications we have in mind — a high-volume, fast-moving supply and demand context, which requires an understanding of the asymptotic behavior of key system performance measures under various supply strategies, as well the pre-planning of such strategies.

As a specific business analytics instance of such capacity problems often encountered within the context of workforce management for large services providers, we note that any significant workforce actions aimed at altering a planned resource supply usually require long lead times, whereas customer demand can exhibit volatility and uncertainty over relatively short periods of time. Competition within the marketplace and the dynamics within these companies create constant pressure for growth; under this pressure, workforce capacity planning routinely examines resource requirements when the demand forecast is scaled up in different proportions. These considerations in part motivate the current problem setting. Another application of a similar flavor concerns capacity provisioning in cloud computing environments where high-frequency demand having considerable uncertainty needs to be fulfilled with a predetermined supply of resources. Any changes in these supply decisions are costly and should be avoided.

Now, let S_n denote the net demand at period $n = 1, \ldots, N$, namely the difference between demand and supply (both in terms of the cumulative quantities) up to period n. In the backorder model, where unmet demand in any period will be backlogged and supplied by future supplies, backorder and inventory are, respectively, the positive and negative parts of S_n; as such, the model is quite tractable. On the other hand, the lost-sales model, in which unmet demand in any period will be lost, is much harder to solve because the cumulative lost-sales quantity, L_n, is the running maximum of S_n: $L_n = \max_{0 \le j \le n}\{S_j\}$ (with $S_0 := 0$), which depends on $S_1, S_2, \ldots, S_n$. When supply is linear, i.e., $\alpha = 1$, then S_n is a random walk and L_n is a well-studied object; in particular, $\mathsf{E}[L_n]$

is well-known and very accessible. For other values of α, however, $\mathsf{E}[L_n]$ does not appear to have explicit expressions. The only possible exception is $\alpha = \frac{1}{2}$, in which case one can approximate S_n by a Brownian motion with square-root drift. We derive such an analysis by relating this Brownian motion to the hitting time of an Ornstein-Uhlenbeck process, for which the density function can be obtained by inverting Laplace transforms.

Given this level of difficulty, our approach consists of constructing upper- and lower-bounds for $\mathsf{E}[L_N]$, and analyzing the asymptotic behavior of these bounds for large N. Our findings for such asymptotic behaviors are summarized as follows. When $\kappa > 0$ (the over-supply case), setting α within $(\frac{1}{2}, 1]$ will result in a bounded $\mathsf{E}[L_N]$, whereas $\alpha \in [0, \frac{1}{2}]$ will cause $\mathsf{E}[L_N]$ to grow on the order of $\sqrt{N}$. When $\kappa < 0$ (the under-supply case), $\mathsf{E}[L_N]$ will grow on the order of N^α upon setting α within $(\frac{1}{2}, 1]$; whereas it will grow on the order of $\sqrt{N}$ when $\alpha \in [0, \frac{1}{2}]$.

In terms of minimizing the total cost over both production cost and lost-sales penalty, our results imply that setting $\alpha = \frac{1}{2}$ is the only meaningful choice leading to non-trivial results, independent of the values of the penalty and holding costs. Hence, what remains is to find the best value of κ. We first obtain an asymptotically optimal solution through a Brownian approximation of the objective function for large N. Then, for each finite N, we examine optimal solutions through the upper and lower bounds of $\mathsf{E}[L_N]$ as surrogates for the original problem. Both bounds lead to simple optimal solutions, and our results include performance guarantees for both problems as approximations to the original problem. We also show that these results are readily extended to include holding costs. In addition, we establish the equivalence between the lost-sales model and the backorder model when both have the same penalty cost that goes to infinity.

1.1. Related Work

A brief review of the related literature is in order. To the extent that our set of supply decisions has to be made before the season starts, our model resembles the newsvendor model. There are important distinctions, however. To start with, the newsvendor decision is a stocking quantity decision, whereas we are concerned with the supply *dynamics* over time — all N periods of the planning horizon. Consequently, the lost-sales quantity in the newsvendor model is a single number that can be accounted for at the end of the horizon. In our model, lost sales build up over the horizon, as demand and supply are realized and evolve. Granted, under our cost model which minimizes the total production cost and lost-sales penalty over the horizon, the lost-sales quantity $\mathsf{E}[L_N]$ is associated with the end of the horizon, as in the newsvendor model. However, we consider inventory holding costs as well the backorder alternative to lost sales, features that are not present in the newsvendor model.

The lost sales problem is known to have a much more complicated structure than the problem of backorder, and hence it is more difficult to analyze. Regarding the optimal policy, the exact

description is not known even in the simplest case. Karlin and Scarf [12] demonstrate that the base-stock policies are not optimal even for systems with a lead time of one period. Monotonicity properties of the optimal ordering policy are derived by Morton [15], Zipkin [23] and most recently Huh [11], each through different methods. Huh *et al.* [10] show that, asymptotically as the ratio of unit penalty cost and unit holding cost grows to infinity, the difference in performance between a particular base-stock policy and the optimal policy will eventually vanish. Thus, base-stock policies can be regarded as asymptotically optimal when there is such a significant imbalance between the penalty cost and holding cost.

Levi *et al.* [14] study production and lost sales over multiple periods focusing on the performance of a simple heuristic policy that balances the inventory holding cost and lost-sales penalty in each period, with the aim of establishing its robustness relative to the optimal policy in a very general context. The setting of Levi *et al.* is dynamic programming: in each period a different decision can be made and actions taken accordingly. In contrast, we allow only a single set of decisions at the beginning of the horizon; and all subsequent costs due to production, lost-sales or backlog, and holding inventory are consequences of this set of decisions. The way we quantify our heuristic solution (from the upper-bound problem), however, appears to have a similar flavor to the proven "2-approximation" status of the dual-balancing rule in [14]. Specifically, the performance of our heuristic, via the ratio of the upper- and lower-bound solutions, is "2-plus" — where this slightly worse performance guarantee than a 2-approximation algorithm can be attributed to the fact that there is no recourse over the planning horizon in our model, which, to make matters worse, can be infinitely long. A complementary set of results for the lost-sales model has been recently obtained by Goldberg *et al.* [8].

Another body of research related to our analysis concerns process flexibility; see the recent paper of Chou *et al.* [7] and the references therein. Specifically, process flexibility can be modeled by a random walk with functional drift, called a generalized random walk in [7]. When proper functional forms are taken, it can be seen that the cost structure under full flexibility has the same form as the cost structure of our backorder model, while the cost structure under a so-called chaining option corresponds to the structure of our lost-sales model. Upon applying the methodology developed in this paper, the effectiveness of chaining can be demonstrated and quantified for more general processes.

1.2. Paper Organization

The remainder of the paper is organized as follows. In Section 2, we start with a formulation of our models — while focusing on the lost-sales model, we also cover the backorder model, followed by a derivation of the main performance measures associated with these models. We then present preliminaries regarding the normal distribution function and the loss function, and derive estimates

for $\mathsf{E}[L_N]$ in the case of $\alpha = 1$. In Section 3, we continue with estimates of $\mathsf{E}[L_N]$ for all other cases, through constructing upper- and lower-bounds and also deriving a corresponding Brownian approximation for the case of $\alpha = 1/2$. Optimization models that minimize total cost over production, lost sales (or backorder) and inventory are studied in Section 4, followed by concluding remarks in Section 5.

2. Problem Formulation and Preliminaries

2.1. Key Performance Measures

Consider a planning horizon that consists of N periods, indexed by $n = 1, \ldots, N$. For each period n, let D_n denote the demand, a random variable; and x_n the production or supply quantity, a decision variable. At the beginning of the planning horizon, we need to determine the production quantities for all N periods, as part of the capacity planning problem. Once this set of decisions is made, then in each period n there will be x_n units available along with any units left over from previous periods to supply demand D_n. In the lost-sales model, any demand surpassing the available supply is lost by the end of each period; whereas in the backorder model, any unfilled demand will be backlogged, to be supplied in a later period. Correspondingly, the quantities of interest include the following.

- L_n: the cumulative demand shortfall, in the lost-sales model, up to period n;
- B_n: the backlogged demand, in the backorder model, at the end of each period n;
- H_n and H'_n: the on-hand inventory left at the end of each period n, respectively for the lost-sales and backorder models.

PROPOSITION 1. Define $S_0 := 0$ and $S_n := D_1 - x_1 + \cdots + D_n - x_n$. Further define $x^+ := \max\{x, 0\}$ and $x^- := -\min\{x, 0\}$. Then, for $n = 1, \ldots, N$, we have

$$B_n = S_n^-, \qquad H'_n = S_n^+, \tag{1}$$

$$L_n = \max_{0 \le j \le n}\{S_j\}, \qquad H_n = L_n - S_n. \tag{2}$$

Proof. In the backorder model, any unmet demand (if $S_n > 0$) or any leftover inventory (if $S_n < 0$) at the end of each period n will be carried over; hence, leading to the expressions in (1).

The H_n expression in (2) follows from

$$H_n = \sum_{i=1}^{n} x_i - \left(\sum_{i=1}^{n} D_i - L_n\right),$$

i.e., total production minus total *supplied* demand equals inventory.

For the L_n expression in (2), note that the result holds trivially for $n = 1$: $L_1 = \max\{0, D_1 - x_1\}$. Suppose it holds for n. Then, we have $L_n = S_k$ for some $k \le n$, which means that there will be no

more shortage for periods $k+1$ through n. Therefore, the units available to supply D_{n+1} consist of x_{n+1} and the leftover from periods $k+1$ through n, namely

$$-(D_{k+1} - x_{k+1} + \cdots + D_n - x_n).$$

(Note that the above expression is non-negative, as the quantity in parentheses cannot exceed 0 due to no shortage from periods $k+1$ through n.) Hence, at the end of period $n+1$, shortage will occur if

$$D_{n+1} - [x_{n+1} - (D_{k+1} - x_{k+1} + \cdots + D_n - x_n)] > 0,$$

in which case the cumulative shortage will be adding S_k to the above expression, with the sum being equal to S_{n+1}. Otherwise, i.e., if the left-hand side above is ≤ 0, then the cumulative shortage stays at S_k. In summary, we have

$$L_{n+1} = \max\{S_k, S_{n+1}\} = \max\{L_n, S_{n+1}\} = \max_{0 \leq j \leq n+1}\{S_j\},$$

which is as desired. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

In what follows, we shall focus on L_n, since H_n relates directly to L_n; whereas B_n and H'_n are simply the positive and negative parts of the random walk S_n, and thus both are easily accessible – refer to (11) below.

Since the study of large N asymptotics is our primary objective, we focus on the normal distribution for the demand model. Specifically, let

$$D_n = \mu_n + \sigma Z_n,$$

where Z_n are independent and identically distributed (i.i.d.) standard normal random variables, and μ_n and σ are the mean and standard deviation of D_n, respectively. The common σ can be interpreted as a forecast error; hence, it is independent of the periods. Alternatively, we can allow a period-dependent σ_n, as long as the sum $\sum_{i=1}^{n} \sigma_i \sim \sigma\sqrt{n}$, i.e., it grows at a rate of $\sqrt{n}$. In this case, the above expression will still serve as a good approximation due to the central limit theorem.

On the production side, we write, for $n = 1, \ldots, N$,

$$x_1 + \cdots + x_n = \mu_1 + \cdots + \mu_n + \kappa\sigma n^\alpha, \tag{3}$$

where κ and $\alpha \in [0,1]$ are two policy parameters (or decision variables). Note that the (cumulative) production quantity in (3) consists of two parts: the first part matches the mean of the (cumulative) demand; and the second part can be interpreted as *safety stock* to offset demand variability. The parameter α is often referred to as the "safety factor". This factor addresses the issue of how the safety stock should match up with the demand variability, where the latter grows on the order $\sqrt{n}$. Recall that κ can take on positive or negative values: when it is positive, $\kappa\sigma n^\alpha$ is truly the

safety stock; when κ is negative, this part is a calculated under-production, suitable for settings where the production cost is much higher relative to the lost-sales penalty. Since (3) implies that the production quantity for each period n is given by

$$x_n = \mu_n + \kappa\sigma[n^\alpha - (n-1)^\alpha], \qquad n = 1, \ldots, N, \tag{4}$$

then the cumulative shortage process L_n can be estimated in different ways depending on the value of α.

2.2. Useful Facts about the Standard Normal Variable

We summarize below some useful expressions and estimates associated with the standard normal variable. Let Z denote the standard normal variate with density and distribution functions denoted by $\phi(x)$ and $\Phi(x)$, respectively. Define the "shortfall function" (or "loss function") as

$$G(x) := \mathsf{E}[(Z - x)^+] = \int_x^{+\infty} (z - x)\phi(z)dz = \phi(x) - x\bar{\Phi}(x), \tag{5}$$

where $\bar{\Phi}(x) := 1 - \Phi(x)$ and the last equality follows directly from $\phi'(z) = -z\phi(z)$.

Now, we list properties of $G(x)$ that will be useful later for our purposes.

1. $G(x)$ is decreasing and convex in x (since $(Z - x)^+$ is decreasing and convex in x).

2. The part of $G(x)$ that has significant curvature is limited to the neighborhood of the origin $x = 0$. For large $x > 0$, we have $G(x) \approx 0$ and $G(-x) \approx x$.

3. A direct derivation yields:

$$2\int_a^b G(x)dx = \Phi(b) - \Phi(a) + bG(b) - aG(a). \tag{6}$$

4.

$$0 \leq G(x) \leq \frac{\phi(x)}{x^2}, \qquad \forall x > 0. \tag{7}$$

5.

$$\frac{x^2 G(x)}{\phi(x)} \to 1 \qquad \text{as } x \to +\infty. \tag{8}$$

To verify (7) and (8), observe that

$$\left(1 - \frac{1}{x^2}\right)\frac{1}{x}\phi(x) \leq \bar{\Phi}(x) \leq \frac{1}{x}\phi(x), \qquad x > 0, \tag{9}$$

where the upper bound follows from

$$\int_x^\infty e^{-u^2/2}du \leq \int_x^\infty \left(1 + \frac{1}{u^2}\right)e^{-u^2/2}du = \frac{1}{x}e^{-x^2/2}$$

and the lower bound follows from

$$\int_x^\infty e^{-u^2/2} du \geq \int_x^\infty \left(1 - \frac{3}{u^4}\right) e^{-u^2/2} du = \left(1 - \frac{1}{x^2}\right) \frac{1}{x} e^{-x^2/2}.$$

The (second) inequality in (7) follows immediately from the lower bound in (9). Combining the upper and lower bounds in (9), we have $\bar{\Phi}(x) \approx \frac{1}{x}\phi(x)$ for large (positive) x. This, along with l'Hôpital's rule, helps to verify (8).

2.3. Estimating $\mathsf{E[L_N]}$ for $\alpha = 1$

When $\alpha = 1$ we have, from (4), $x_n = \mu_n + \kappa\sigma$; thus, $D_n - x_n = -\kappa\sigma + \sigma Z_n$. In this case, the partial sums $\{S_n\}$ constitute a random walk, and $\mathsf{E}[L_n]$ follows from standard results concerning the maximum of a random walk. Specifically, by Spitzer's Identity (see [20]; also, Ross [18], Proposition 7.1.5):

$$\mathsf{E}[L_N] = \sum_{n=1}^N \frac{1}{n} \mathsf{E}[S_n^+], \tag{10}$$

where $S_n \stackrel{\mathrm{d}}{=} -\kappa\sigma n + \sigma\sqrt{n}Z$. Hence,

$$\mathsf{E}[S_n^+] = \sigma\sqrt{n}\mathsf{E}[(Z - \kappa\sqrt{n})^+] = \sigma\sqrt{n}G(\kappa\sqrt{n}). \tag{11}$$

Approximating the summation by integration, we have

$$\sum_{n=1}^N \frac{1}{n}\mathsf{E}[S_n^+] \approx \int_0^N \frac{1}{t}\mathsf{E}[S_t^+]dt = \sigma\int_0^N \frac{1}{\sqrt{t}}G(\kappa\sqrt{t})dt = \frac{2\sigma}{\kappa}\int_0^{\kappa\sqrt{N}} G(u)du. \tag{12}$$

Therefore, combining the above with (6), we obtain

$$\mathsf{E}[L_N] \approx \frac{\sigma}{\kappa}\left(\Phi(\kappa\sqrt{N}) - \frac{1}{2}\right) + \sigma\sqrt{N}G(\kappa\sqrt{N}). \tag{13}$$

When $\kappa > 0$, the above is increasing in N. As $N \to \infty$, the second term on the right-hand side vanishes, and the first term converges to $\frac{\sigma}{2\kappa}$. When $\kappa < 0$, the second term on the right-hand side of (13) quickly becomes $\sigma|\kappa|N$ as N grows, and the first term approaches a constant $\frac{\sigma}{2|\kappa|}$. These results then can be summarized in the following proposition.

PROPOSITION 2. *For $\alpha = 1$ and large N, we have*

$$\mathsf{E}[L_N] \approx \frac{\sigma}{2\kappa}, \quad \text{when} \quad \kappa > 0; \qquad \mathsf{E}[L_N] \approx \frac{\sigma}{2|\kappa|} + \sigma|\kappa|N, \quad \text{when} \quad \kappa < 0. \tag{14}$$

Hence, setting $\alpha = 1$, the expected shortage is bounded by a constant, or grows linearly in N, corresponding to a positive or negative κ, respectively.

3. Estimating $\mathsf{E}[\mathbf{L_N}]$ for $\alpha \in [\mathbf{0}, \mathbf{1})$

When $\alpha < 1$, it is generally a difficult problem to calculate the exact value of $\mathsf{E}[L_N]$, and thus we develop bounds to reveal its asymptotic order as $N \to \infty$. We shall initially focus on $\kappa > 0$ in the next two subsections, and then summarize the corresponding results for $\kappa \leq 0$ in the following subsection. Lastly, we consider a Brownian approximation for the case when $\alpha = \frac{1}{2}$.

3.1. Case of $\alpha \in (\frac{1}{2}, 1)$ and $\kappa > 0$

From (2), taking into account $S_0 = 0$, we have

$$L_n = \max_{0 \leq j \leq n} \{S_j\} = \max_{1 \leq j \leq n} \{S_j^+\}, \qquad n = 1, \ldots, N.$$

Jensen's inequality implies

$$\mathsf{E}[L_N] \geq \max_{1 \leq n \leq N} \{\mathsf{E}[S_n^+]\}, \tag{15}$$

where, similar to (11), we obtain

$$\mathsf{E}[S_n^+] = \sigma \sqrt{n} \mathsf{E}[(Z - \kappa n^{\alpha - 1/2})^+] = \sigma \sqrt{n} G(\kappa n^{\alpha - 1/2}). \tag{16}$$

From (7), we know that

$$\sigma \sqrt{n} G(\kappa n^{\alpha - 1/2}) \leq \frac{\sigma \sqrt{n} \phi(\kappa n^{\alpha - 1/2})}{\kappa^2 n^{2\alpha - 1}} := \frac{\sigma \sqrt{n} \phi(u)}{u^2}, \qquad \text{where} \qquad u := \kappa n^{\alpha - 1/2}.$$

Hence, $\mathsf{E}[S_n^+]$ will decrease to 0 (since $\phi(u)$ decreases to 0 exponentially fast as $u \to +\infty$). The values of n that achieve the maximum in the lower bound of (15) must satisfy the following optimality equation (whose solution need not be unique):

$$\frac{1}{2} n^{-1/2} G(y) = n^{1/2} \bar{\Phi}(y) \kappa (\alpha - 1/2) n^{\alpha - 3/2}.$$

This simplifies to

$$G(y) = (2\alpha - 1) y \bar{\Phi}(y) \qquad \text{or} \qquad \phi(y) = 2\alpha y \bar{\Phi}(y), \tag{17}$$

and from (9) we know that $\phi(y) \approx y \bar{\Phi}(y)$ when y is large. Since $\alpha > \frac{1}{2}$, the solution to the above equation exists. Indeed, from (9), we have

$$y \bar{\Phi}(y) \geq \left(1 - \frac{1}{y^2}\right) \phi(y), \qquad y > 0.$$

If we replace $y \bar{\Phi}(y)$ in (17) by this lower bound, then the resulting solution

$$\bar{y} := \left(\frac{1}{2\alpha - 1}\right)^{\frac{1}{2}} \tag{18}$$

is an upper bound to the solution of the equations in (17), since $\phi(y)$ is decreasing in y for $y > 0$. Therefore, in this case we obtain

$$\mathsf{E}[L_N] \geq \max_{0 \leq n \leq N}\{\mathsf{E}[S_n^+]\} \approx \sigma\left(\frac{y}{\kappa}\right)^{1/(2\alpha-1)} G(y), \tag{19}$$

where y is the solution to (17) and we know $y \in (0, \bar{y})$.

For an upper bound, we have $L_N = \max_{0 \leq n \leq N}\{S_n\} \leq \sum_{n=1}^{N} S_n^+$, which, along with (16), leads to

$$\mathsf{E}[L_N] \leq \sigma \sum_{n=1}^{N} \sqrt{n} G(\kappa n^{\alpha - \frac{1}{2}}) \approx \sigma \int_0^N \sqrt{t} G(\kappa t^{\alpha - \frac{1}{2}}) dt. \tag{20}$$

Using a transformation of variable $u = t^{\alpha - \frac{1}{2}}$, together with $G(x) \leq \phi(x)/x^2$ from (7), we obtain

$$
\begin{aligned}
\int_0^N \sqrt{t} G(\kappa t^{\alpha - \frac{1}{2}}) dt &= \frac{2}{2\alpha - 1} \int_0^{N^{\alpha - \frac{1}{2}}} u^{\frac{4-2\alpha}{2\alpha-1}} G(\kappa u) du \\
&\leq \frac{2}{(2\alpha - 1)\kappa^2} \int_0^{N^{\alpha - \frac{1}{2}}} u^{\frac{6-6\alpha}{2\alpha-1}} \phi(\kappa u) du \\
&= \frac{2}{2\alpha - 1} \kappa^{-\frac{3}{2\alpha-1}} \int_0^{\kappa N^{\alpha - \frac{1}{2}}} v^{\frac{6-6\alpha}{2\alpha-1}} \phi(v) dv \\
&\leq \frac{2}{2\alpha - 1} \kappa^{-\frac{3}{2\alpha-1}} \int_0^{\infty} v^{\frac{6-6\alpha}{2\alpha-1}} \phi(v) dv := C_\alpha < \infty.
\end{aligned}
\tag{21}
$$

Since $\frac{6-6\alpha}{2\alpha-1} > 0$, as $\alpha \in (\frac{1}{2}, 1)$, we can summarize our results for this case as follows.

PROPOSITION 3. For $1/2 < \alpha < 1$ and $\kappa > 0$, we have

(a) the lower bound:

$$\mathsf{E}[L_N] \geq \sigma\left(\frac{y_\alpha^*}{\kappa}\right)^{1/(2\alpha-1)} G(y_\alpha^*), \quad \text{for} \quad N \geq \bar{y}_\alpha := (2\alpha - 1)^{-1/2}, \tag{22}$$

where $y_\alpha^* > 0$ is the solution to the equation $\phi(y) = 2\alpha y \bar{\Phi}(y)$ (and hence, independent of N);

and

(b) the upper bound: $\mathsf{E}[L_N] \leq \sigma C_\alpha$, with C_α as specified in (21).

Hence, in this case $\mathsf{E}[L_N]$ is both upper- and lower-bounded by constants.

3.2. Case of $\alpha \in [0, \frac{1}{2}]$ and $\kappa > 0$

For $\alpha \in [0, \frac{1}{2}]$, as $n \to \infty$, we have $G(\kappa n^{\alpha - 1/2}) \to G(0)$ or $G(\kappa)$ depending on whether $\alpha < \frac{1}{2}$ or $\alpha = \frac{1}{2}$. Hence, from (15) and (16), we obtain, for sufficiently large N,

$$\mathsf{E}[L_N] \geq \sigma G(\kappa)\sqrt{N}, \quad \text{when } \alpha = \frac{1}{2}; \qquad \mathsf{E}[L_N] \geq \sigma G(0)\sqrt{N}, \quad \text{when } \alpha \in [0, \frac{1}{2}). \tag{23}$$

For an upper bound, observe that the concave function $f(n) = n^\alpha$ is bounded from *below* by a linear function:

$$n^\alpha \geq \frac{N^\alpha}{N} n := \Theta n, \qquad \text{for} \quad 0 \leq n \leq N. \tag{24}$$

Defining $W_n := -\kappa\sigma\Theta n + \sigma\sum_{k=1}^n Z_n$, we know that $L_N \leq \max_{n\leq N} W_n$. Meanwhile, $\max_{n\leq N} W_n$ can be calculated through Spitzer's identity. More specifically, we have

$$\mathsf{E}[W_n^+] = \sigma\sqrt{n}G(\kappa\Theta\sqrt{n}), \qquad 0 \leq n \leq N,$$

and

$$\sum_{n=1}^N \frac{1}{n}\mathsf{E}[W_n^+] \approx \int_0^N \frac{1}{t}\mathsf{E}[W_t^+]dt = \frac{2\sigma}{\kappa\Theta}\int_0^{\kappa\Theta\sqrt{N}} G(x)dx.$$

Therefore, making use of (6), we obtain

$$\mathsf{E}[L_N] \leq \sum_{n=1}^N \frac{1}{n}\mathsf{E}[W_n^+] \approx \frac{\sigma}{\kappa\Theta}\left(\Phi(\kappa\Theta\sqrt{N}) - \frac{1}{2}\right) + \sigma\sqrt{N}G(\kappa\Theta\sqrt{N}). \tag{25}$$

As $N \to \infty$, and according to the definition of Θ in (24), we have $\Theta\sqrt{N} \to 0$ if $\alpha \in [0, \frac{1}{2})$ and $\Theta\sqrt{N} \to 1$ if $\alpha = \frac{1}{2}$. The second term on the right-hand side of (25) is of order $\sqrt{N}$, whereas the first term can be written as $\frac{\sigma\sqrt{N}}{\kappa\Theta\sqrt{N}}[\Phi(\kappa\Theta\sqrt{N}) - \frac{1}{2}]$. Since $\lim_{y\to 0}(\Phi(y) - \frac{1}{2})/y = \phi(0)$, we know this term is on the order of $\sqrt{N}$ as well. Hence,

$$\mathsf{E}[L_N] \leq \sigma\sqrt{N}[\phi(\kappa\Theta\sqrt{N}) + G(\kappa\Theta\sqrt{N})] \leq \sigma\sqrt{N}[\phi(0) + G(0)] = 2\sigma\phi(0)\sqrt{N}. \tag{26}$$

These results allow us to conclude as follows.

PROPOSITION 4. *In the case of $\alpha \in [0, \frac{1}{2}]$ and $\kappa > 0$, $\mathsf{E}[L_N]$ grows on the order of $\sqrt{N}$. In particular, we have the lower- and upper-bounds for $\mathsf{E}[L_N]$ in (23) and (26).*

3.3. Case of $\kappa \leq 0$

First consider the lower bounds. For $\alpha \in (\frac{1}{2}, 1)$ and $\kappa < 0$, we have

$$\mathsf{E}[S_n^+] = \sigma\sqrt{n}G(\kappa n^{\alpha-1/2}) \approx \sigma|\kappa|n^\alpha,$$

since $G(-x) \approx x$ when $x > 0$ is large. Hence, for N sufficiently large,

$$\mathsf{E}[L_N] \geq \max_{0\leq n\leq N}\{\mathsf{E}[S_n^+]\} \approx \sigma|\kappa|N^\alpha, \quad \text{for } \alpha \in (\frac{1}{2}, 1). \tag{27}$$

When $\alpha \in [0, \frac{1}{2}]$ and $\kappa < 0$, the lower bound for $\kappa > 0$ in (23) remains valid; namely,

$$\mathsf{E}[L_N] \geq \sigma G(\kappa)\sqrt{N}, \quad \text{when } \alpha = \frac{1}{2}; \qquad \mathsf{E}[L_N] \geq \sigma G(0)\sqrt{N}, \quad \text{when } \alpha \in [0, \frac{1}{2}). \tag{28}$$

Next, we consider the upper bounds. Define $Y_n := S_n + \kappa\sigma n^\alpha$, for $n = 0, 1, \ldots, N$. Note that Y_n is a random walk with zero drift and $Y_0 = 0$. Then,

$$L_N = \max_{0\leq n\leq N}\{S_n\} = \max_{0\leq n\leq N}\{Y_n - \kappa\sigma n^\alpha\}$$

$$\leq \max_{0 \leq n \leq N} \{Y_n\} + \max_{0 \leq n \leq N} \{-\kappa \sigma n^\alpha\}$$

$$= \max_{0 \leq n \leq N} \{Y_n\} + (-\kappa \sigma N^\alpha), \tag{29}$$

where the inequality follows from the subadditivity of the maximum operator, and the last equality is due to $\kappa < 0$. Applying Spitzer's identity to the random walk Y_n, similar to the derivation of $\mathsf{E}[L_N]$ in (13) for the case of $\alpha = 1$, we obtain

$$\mathsf{E}\left[\max_{0 \leq n \leq N} \{Y_n\}\right] = \mathsf{E}\left[\max_{1 \leq n \leq N} \{Y_n^+\}\right] = \sum_{n=1}^{N} \frac{\sigma \sqrt{n}}{n} G(0) \approx \sigma G(0) \int_0^N \frac{dt}{\sqrt{t}} = 2\sigma G(0)\sqrt{N}.$$

Hence, combining this with (29), renders

$$\mathsf{E}[L_N] \leq 2\sigma G(0)\sqrt{N} + |\kappa|\sigma N^\alpha, \qquad \alpha \geq 0. \tag{30}$$

Our results for this case then can be summarized in the following proposition.

PROPOSITION 5. *For $\kappa < 0$, $\mathsf{E}[L_N]$ grows on the order of $\sigma|\kappa|N^\alpha$ when $\alpha \in (\frac{1}{2}, 1)$; it grows on the order of $\sigma G(0)\sqrt{N}$ when $\alpha \in [0, \frac{1}{2})$; and it grows on the order of $\sigma G(k)\sqrt{N}$ when $\alpha = \frac{1}{2}$. In particular, the lower bound of $\mathsf{E}[L_N]$ follows (27) and (28), with its upper bound following (30).*

Finally, when $\kappa = 0$, we have from (29) and (30)

$$\mathsf{E}[L_N] = \mathsf{E}\left[\max_{0 \leq n \leq N} \{Y_n\}\right] = 2\sigma G(0)\sqrt{N},$$

which is also consistent with setting $\kappa = 0$ in the lower bound (28) and in the upper bound (30).

3.4. Brownian approximation for case of $\alpha = \frac{1}{2}$

The case of $\alpha = \frac{1}{2}$ will turn out to produce the right trade-off between cost and service, as we will see in the next section, thus yielding the optimal order of production. It is therefore appropriate to explore the possibility of a more accurate estimation of L_n in this particular case.

To this end, we study the continuous counterpart of L_n, i.e., a Brownian motion with square-root drift. More specifically, $\mathsf{E}[L_N]$ can be approximated as

$$\mathsf{E}[L_N] \approx \sqrt{\sigma} \mathsf{E}\left[\sup_{1 \leq s \leq t} \left\{B(s) - \frac{\kappa}{\sqrt{\sigma}}\sqrt{s}\right\}\right], \tag{31}$$

where $B(t)$ is a standard Brownian motion. Defining

$$\tau_x := \inf\left\{t > 1 : B(t) = \frac{\kappa}{\sqrt{\sigma}}\sqrt{t}; B(1) = -x\right\}$$

for any $x \geq 0$, we then have

$$\mathsf{E}\left[\sup_{1 \leq s \leq t} \left\{B(s) - \frac{\kappa}{\sqrt{\sigma}}\sqrt{s}\right\}\right] = \int_0^\infty \mathsf{P}\left[\sup_{1 \leq s \leq t} \left\{B(s) - \frac{\kappa}{\sqrt{\sigma}}\sqrt{s}\right\} \geq x\right] dx$$

$$= \int_0^\infty \mathsf{P}\left[\tau_x < t\right] dx. \tag{32}$$

Now, define a new process $Y(u) := B(e^{2u})/e^u$. One can readily verify, as shown in [4], that $Y(u)$ is an Ornstein-Uhlenbeck process with $Y(0) = x$. Further define

$$T_{a,b} := \inf\{u \geq 0 : Y(u) \geq b, Y(0) = a\}.$$

We note that, for any $t \geq 0$,

$$\mathsf{P}[\tau_x > e^{2t}] = \mathsf{P}[T_{-x,\kappa/\sqrt{\sigma}} > t], \tag{33}$$

a fact that has been known since [4]. Our task is thus reduced to the evaluation of $T_{a,b}$. When $b = 0$, an explicit expression for $T_{a,b}$ has been derived; see, e.g., [17, 13, 9]. More generally when $b \neq 0$ (as is of interest here), however, only the Laplace transform of the density of $T_{a,b}$ is known. Using various methods, the moments of this distribution also have been calculated; refer to, e.g., [22, 21, 5, 16, 17]. In theory, the distribution function of $T_{a,b}$ can be recovered from all of its moments. In practice, the distribution function of $T_{a,b}$ can be effectively approximated by its finite moments up to a certain order. Moreover, as shown in [3], the error bound of such an approximation can be computed through numerical methods, such as semi-definite programming.

We therefore propose to use the first K moments $M_{a,b}^1, \ldots, M_{a,b}^K$ of $T_{a,b}$ where $a = -x$ and $b = \kappa/\sqrt{\sigma}$, in combination with (31), (32) and (33), to obtain a more accurate estimation of L_n in the case of $\alpha = \frac{1}{2}$. The methods and results in [22, 16, 17] can be used to calculate the first K moments $M_{a,b}^1, \ldots, M_{a,b}^K$. When $K \leq 3$, we can additionally exploit the closed-form tight bounds in [3] to further improve our approximation of the distribution function of $T_{a,b}$. (Finding the best possible bounds for $K \geq 4$ is NP-hard [3].) As a specific instance of our proposed approach, we consider in more detail the case of utilizing the first three moments of $T_{a,b}$ (i.e., $K = 3$). Our starting point is the set of closed-form expressions for $M_{a,b}^1, M_{a,b}^2, M_{a,b}^3$ provided in [16] and parameterized by $a = -x$ and $b = \kappa/\sqrt{\sigma}$. Then, letting X denote a generic random variable that follows the probability distribution function of $T_{-x,\kappa/\sqrt{\sigma}}$, we next obtain closed-form tight upper bounds on both $\mathsf{P}[X \leq z]$ and $\mathsf{P}[X \geq z]$ in terms of the expressions for $M_{-x,\kappa/\sqrt{\sigma}}^1, M_{-x,\kappa/\sqrt{\sigma}}^2, M_{-x,\kappa/\sqrt{\sigma}}^3$. More specifically, we exploit the following propostion adapted from Theorem 3.3 in [3].

PROPOSITION 6. *For a constant $\delta > 0$ and a nonnegative real random variable $X \stackrel{\mathrm{d}}{=} T_{-x,\kappa/\sqrt{\sigma}}$ with first three moments $M_1 = M_{-x,\kappa/\sqrt{\sigma}}^1$, $M_2 = M_{-x,\kappa/\sqrt{\sigma}}^2$ and $M_3 = M_{-x,\kappa/\sqrt{\sigma}}^3$, the functionals $\mathsf{P}[X > (1+\delta)M_1]$ and $\mathsf{P}[X < (1-\delta)M_1]$ can be approximated as*

$$\mathsf{P}[X > (1+\delta)M_1] \approx f_1(C_M^2, D_M^2, \delta) \qquad \text{and} \qquad \mathsf{P}[X < (1-\delta)M_1] \approx f_2(C_M^2, D_M^2, \delta),$$

respectively, where

$$f_1(C_M^2, D_M^2, \delta) = \begin{cases} \min\left(\frac{C_M^2}{C_M^2 + \delta^2}, \frac{1}{1+\delta} \cdot \frac{D_M^2}{D_M^2 + (C_M^2 - \delta)^2}\right), & \delta > C_M^2, \\ \frac{1}{1+\delta} \cdot \frac{D_M^2 + (1+\delta)(C_M^2 - \delta)}{D_M^2 + (1+C_M^2)(C_M^2 - \delta)}, & \delta \leq C_M^2, \end{cases}$$

$$f_2(C_M^2, D_M^2, \delta) = 1 - \frac{(C_M^2 + \delta)^3}{(D_M^2 + (C_M^2 + 1)(C_M^2 + \delta))(D_M^2 + (C_M^2 + \delta)^2)}, \qquad \delta < 1,$$

$$C_M^2 = \frac{M_2 - M_1^2}{M_1^2},$$

$$D_M^2 = \frac{M_1 M_3 - M_2^2}{M_1^4}.$$

Lastly, we combine Proposition 6 and the expressions for $M^1_{-x,\kappa/\sqrt{\sigma}}, M^2_{-x,\kappa/\sqrt{\sigma}}, M^3_{-x,\kappa/\sqrt{\sigma}}$ in [16] together with (31), (32), (33) to obtain the more accurate approximation of L_n in the case of $\alpha = \frac{1}{2}$, as desired.

4. Cost Minimization and Asymptotically Optimal Solutions

In this section we consider instances of two cost minimization models for the lost-sale system of interest. For the first model, the total cost consists of both the production cost and lost-sale penalty. To be more precise, we minimize $c \sum_{n=1}^N x_n + p\mathsf{E}[L_N]$ where c and p are the per-unit production cost and lost-sales penalty, respectively. From (3), ignoring the constant term $c(\mu_1 + \cdots + \mu_N)$, we have

$$\min_{\alpha, \kappa} \quad c\sigma\kappa N^\alpha + p\mathsf{E}[L_N]. \tag{34}$$

For the second model, we incorporate the inventory (holding) cost in each period. Specifically, we have

$$\min_{\alpha, \kappa} \quad c\sigma\kappa N^\alpha + p\mathsf{E}[L_N] + h\mathsf{E}[H_N], \tag{35}$$

where h is the per-unit inventory holding cost. Recall from Proposition 1 that the inventory (at the end) of period n is given by $H_n = L_n - S_n$, and therefore $\mathsf{E}[H_N] = \mathsf{E}[L_N] + \kappa\sigma N^\alpha$.

Our first step is to find the optimal value of α in these optimization problems. For this purpose, let us start with a summary of the results we have derived so far concerning the asymptotics of the expected loss sales in the limit as $N \to \infty$.

(i) When $\alpha = 1$, $\mathsf{E}[L_N]$ is bounded by a constant $\frac{\sigma}{2\kappa}$ if $\kappa > 0$; whereas it grows linearly in N if $\kappa < 0$.

(ii) When $\alpha \in (\frac{1}{2}, 1)$ and $\kappa > 0$, $\mathsf{E}[L_N]$ is bounded from above and below by constants that are independent of N (but dependent on α and κ); whereas it grows on the order of N^α if $\kappa < 0$.

(iii) When $\alpha \in [0, \frac{1}{2}]$, $\mathsf{E}[L_N]$ grows on the order of $\sqrt{N}$, regardless of whether $\kappa > 0$ or $\kappa \le 0$.

From (34), if $\alpha \in (\frac{1}{2}, 1]$ and $\kappa > 0$, then the second term in the objective will be bounded by a constant independent of N; hence, for sufficiently large N, the objective is of order N^α, i.e., the same order as the first term. On the other hand, if $\alpha \in (\frac{1}{2}, 1]$ and $\kappa < 0$, then the first term will decrease while the second term will increase, both on the order of $|\kappa|N^\alpha$; hence, if $c > p$, then $\kappa = -\infty$ minimizes the objective value, whereas if $c \le p$, then $\kappa \to 0$ minimizes the objective value.

Therefore, the only non-trivial solution of the optimization problem is to set $\alpha \leq \frac{1}{2}$, corresponding to the case summarized in (iii) above. Here, the first term of the objective function is of order N^{α}, which is dominated by the second term of order $\sqrt{N}$ (except when $\alpha = \frac{1}{2}$). Consequently, the objective value will also be of order $\sqrt{N}$, as is further confirmed by replacing $\mathsf{E}[L_N]$ with its lower bound via Jensen's inequality (which applies in all cases): $\mathsf{E}[L_N] \geq \mathsf{E}[\max_{1 \leq n \leq N} S_n^+] \geq \mathsf{E}[S_N^+]$. This leads to the following minimization problem, which is a lower bound of the original problem:

$$\min_{z} \quad cz + p\mathsf{E}\left[\left(\sigma \sum_{n=1}^{N} Z_n - z\right)^{+}\right] = cz + p\mathsf{E}\left[(\sigma\sqrt{N}Z - z)^{+}\right]. \tag{36}$$

To allow for any $\alpha \in [0,1]$ and any κ, we write

$$z = \sigma\kappa N^{\alpha} = \sigma\sqrt{N}\kappa N^{\alpha-\frac{1}{2}} := \sigma\sqrt{N}y,$$

and then the optimal solution to minimizing $cy + p\mathsf{E}[(Z - y)^{+}]$ is given by

$$y^* = \bar{\Phi}^{-1}\left(\frac{c}{p}\right) = \Phi^{-1}\left(\frac{p-c}{p}\right). \tag{37}$$

With y^* being a constant, independent of N, we must have that $\alpha = \frac{1}{2}$, and hence $\kappa = y^*$. The corresponding objective value then can be expressed as

$$\sigma\sqrt{N}[cy^* + pG(y^*)] = \sigma\sqrt{N}[cy^* + p\phi(y^*) - py^*\bar{\Phi}(y^*)] = \sigma\sqrt{N}p\phi(y^*). \tag{38}$$

This confirms our previous statements; namely, the minimal overall cost cannot be lower than order $\sqrt{N}$. Moreover, we shall henceforth assume that $p \geq c$, unless noted otherwise, because if $p < c$ then the lower-bound solution in (37) already indicates what will happen: make y as small (negative) as possible; whereas the objective value is lower bounded by the expression in (38). We therefore only need to consider the case of $\alpha = \frac{1}{2}$. Furthermore, upon careful examination of the objective of the second optimization model, we can reach the same conclusion.

In the remainder of this section, we start by considering asymptotically optimal solutions of the above optimization models using two different approaches to obtain the desired solutions. We then turn to incorporate inventory costs in our optimization models. Finally, we consider the asymptotic equivalence of lost-sales and backorder models under certain conditions.

4.1. Asymptotically Optimal Solutions

With our focus on $\alpha = \frac{1}{2}$, we first obtain an asymptotically optimal solution through a Brownian approximation of the objective function for large N. We then obtain optimal solutions through an analysis of the asymptotic behavior of our upper and lower bounds of $\mathsf{E}[L_N]$.

4.1.1. Brownian Approximation   Recall that the optimization problem for the production loss trade-off, when α is fixed to be $\frac{1}{2}$, has the form:

$$\min_{\kappa} \quad c\sigma\kappa N^{1/2} + p\mathsf{E}[L_N]. \tag{39}$$

Taking a closer look at the objective function, we conclude

$$c\sigma\kappa N^{1/2} + p\mathsf{E}[L_N] = \left(c\sigma\kappa + p\frac{\mathsf{E}[L_N]}{N^{1/2}} \right) N^{1/2}.$$

Moreover,

$$\frac{\mathsf{E}[L_N]}{N^{1/2}} = \mathsf{E}\left[\max_{n \leq N} \frac{S_n}{N^{1/2}} \right] = \mathsf{E}\left[\max_{n \leq N} \frac{\sum_{i=1}^{n} \sigma Z_i - \kappa\sigma\sqrt{n}}{N^{1/2}} \right]$$

where Z_i are i.i.d. copies of a random variable following the standard normal distribution. This leads to our next result.

THEOREM 7.   Let $\rho(\kappa) = \sup_{t \leq 1} B_t - \kappa\sqrt{t}$, where B_t is a standard Brown motion. Then,

$$\lim_{N \to \infty} \mathsf{E}\left[\max_{n \leq N} \frac{\sum_{i=1}^{n} Z_i - \kappa\sqrt{n}}{N^{1/2}} \right] \to \rho(\kappa).$$

Proof.   We first obtain

$$\max_{n \leq N} \frac{\sum_{i=1}^{n} Z_i - \kappa\sqrt{n}}{N^{1/2}} = \max_{t \in \{\frac{n}{N}, n=1,2,\dots,N\}} \frac{\sum_{i=1}^{Nt} Z_i - \kappa\sqrt{Nt}}{N^{1/2}}$$

$$= \sup_{t \in (0,1]} \frac{\sum_{i=1}^{\lfloor Nt \rfloor} Z_i - \kappa\sqrt{\lfloor Nt \rfloor}}{N^{1/2}}.$$

From the functional central limit theorem, for the summation of random variables (see e.g., [6]), we can conclude that

$$\frac{\sum_{i=1}^{\lfloor Nt \rfloor} Z_i - \kappa\sqrt{\lfloor Nt \rfloor}}{N^{1/2}} \Rightarrow B_t - \sqrt{t}.$$

The desired convergence then follows from the continuous mapping theorem. $\square$

The above theorem implies that, for any $\epsilon > 0$ and when N is large enough, we have

$$\left| \frac{\mathsf{E}[L_N]}{N^{1/2}} - \rho(\kappa) \right| < \epsilon$$

uniformly on compact sets in κ. Although there is no explicit formula for $\rho(\kappa)$, computational methods that make use of the related results in [2, 4, 17, 21] together with Monte-Carlo simulation can be employed; in particular, these methods can be used to obtain accurate estimation of the value of $\rho(\kappa)$ and its derivative in order to solve the Brownian version of our optimization problem:

$$\min_{\kappa} \quad c\sigma\kappa N^{1/2} + p\sigma\rho(\kappa). \tag{40}$$

Let κ^* denote an optimum of the problem (40); of course, κ^* corresponds to a specific production plan. We next show that this production plan is asymptotically optimal, defined as follows.

DEFINITION 8. A production plan is *asymptotically optimal* if

$$\limsup_{N \to \infty} \frac{J_N(\kappa)}{J_N^*} \leq 1.$$

THEOREM 9. The optimal solution of (40), κ^*, is asymptotically optimal for the problem (39).

Proof. Let $\kappa^*(N)$ denote an optimum of problem (39) for a given N. Then, for any N, we have

$$\frac{J_N(\kappa)}{J_N^*} = \frac{c\sigma\kappa^* N^{1/2} + p\mathsf{E}[L_N(\kappa^*)]}{c\sigma\kappa^*(N)N^{1/2} + p\mathsf{E}[L_N(\kappa^*(N))]}$$

$$= \frac{c\sigma\kappa^* N^{1/2} + p\mathsf{E}[L_N(\kappa^*)] - \left(c\sigma\kappa^*(N)N^{1/2} + p\mathsf{E}[L_N(\kappa^*(N))]\right) + \left(c\sigma\kappa^*(N)N^{1/2} + p\mathsf{E}[L_N(\kappa^*(N))]\right)}{c\sigma\kappa^*(N)N^{1/2} + p\mathsf{E}[L_N(\kappa^*(N))]}$$

$$\leq \frac{c\sigma\kappa^* N^{1/2} + p\mathsf{E}[L_N(\kappa^*)] - \left(c\sigma\kappa^*(N)N^{1/2} + p\mathsf{E}[L_N(\kappa^*(N))]\right) +}{c\sigma\kappa^*(N)N^{1/2} + p\mathsf{E}[L_N(\kappa^*(N))]} + 1.$$

We further know that the first term can be bounded by a constant, and thus goes to 0 in the limit at $N \to \infty$. The desired result of asymptotic optimality follows. □

It is evident that the asymptotically optimal result is not restricted to the production optimization problem, as a similar conclusion can be also reached for the inventory optimization problem.

4.1.2. Lower- and Upper-Bound Surrogate Problems We now turn to consider an alternative approach for obtaining optimal solutions through our upper and lower bounds of $\mathsf{E}[L_N]$. For finite $N > 0$, we have already analyzed the lower bound surrogate problem in (38). We next seek to minimize the upper bound with respect to the choice of κ. From the upper bound for the case $\alpha = \frac{1}{2}$ and $\kappa > 0$ in terms of (24) and (25), upon replacing $c\sigma\sqrt{N}$ and $p\sigma\sqrt{N}$ with c and p, respectively, we want to solve the following optimization problem:

$$\min_{\kappa \geq 0} c\kappa + p\left(\frac{1}{\kappa}\left[\frac{1}{2} - \bar{\Phi}(\kappa)\right] + G(\kappa)\right). \tag{41}$$

The optimality condition is then given by

$$c + \frac{p}{\kappa^2}\left(\kappa\phi(\kappa) + \bar{\Phi}(\kappa) - \frac{1}{2}\right) = p\bar{\Phi}(\kappa),$$

which simplifies to

$$c\kappa + pG(\kappa) = \frac{p}{\kappa}\left(\frac{1}{2} - \bar{\Phi}(\kappa)\right). \tag{42}$$

Denoting the solution of (42) by κ_u and observing $\kappa_u \geq 0$ implies that $\bar{\Phi}(\kappa_u) \leq \frac{1}{2}$. Further denote the objective values of the lower- and upper-bound problems in (38) and (41) by V^ℓ and V^u, respectively. We write $\kappa_\ell := y^*$ following (37), which yields

$$\kappa_\ell := \bar{\Phi}^{-1}\left(\frac{c}{p}\right) = \Phi^{-1}\left(\frac{p-c}{p}\right). \tag{43}$$

Then, from (38), (41) and (42), we have

$$\frac{V^u}{V^\ell} = \frac{2[c\kappa_u + pG(\kappa_u)]}{c\kappa_\ell + pG(\kappa_\ell)} \geq 2, \tag{44}$$

since κ_ℓ is the minimizer of $cy + pG(y)$.

To go in the opposite direction, from (38), (41) and (42), we can also write the above ratio as

$$\frac{V^u}{V^\ell} = \frac{2[\frac{1}{2} - \bar{\Phi}(\kappa_u)]}{\kappa_u \phi(\kappa_\ell)}.$$

Observe that $\frac{1}{x}[\frac{1}{2} - \bar{\Phi}(x)]$ is decreasing in $x \geq 0$ (and at $x = 0$, the limit is $\phi(0)$, via l'Hôpital's rule), since its derivative is $\frac{1}{x^2}[x\phi(x) + \bar{\Phi}(x) - \frac{1}{2}]$. It is then straightforward to verify that $x\phi(x) + \bar{\Phi}(x) - \frac{1}{2} \leq 0$ for $x \geq 0$. We therefore have

$$\phi(0) \geq \frac{1}{\kappa_u}\left(\frac{1}{2} - \bar{\Phi}(\kappa_u)\right),$$

and hence

$$\frac{V^u}{V^\ell} \leq \frac{2\phi(0)}{\phi(\kappa_\ell)}. \tag{45}$$

In fact, we also know that $\kappa_\ell \leq \kappa_u$, which follows from

$$\frac{p}{\kappa_\ell}\left(\frac{1}{2} - \bar{\Phi}(\kappa_\ell)\right) \geq \frac{p}{\kappa_u}\left(\frac{1}{2} - \bar{\Phi}(\kappa_u)\right)$$

(since κ_u is the minimizer of the right-hand side), together with the fact justified above that $\frac{1}{x}[\frac{1}{2} - \bar{\Phi}(x)]$ is decreasing in x. When $p = 2c$, we have $\kappa_\ell = 0$; refer to (43). Then, we will have $\kappa_u = 0$ as well; refer to (42). These results render, for this case, $V^u = 2V^\ell$.

In summary, we have established the following proposition.

PROPOSITION 10. *Let V^* denote the optimal value of the original problem in (34). Let V^ℓ and V^u denote the objective values of the lower- and upper-bound problems in (38) and (41), respectively. Then,*

$$\frac{V^u}{V^\ell} = \frac{2[c\kappa_u + pG(\kappa_u)]}{c\kappa_\ell + pG(\kappa_\ell)} = \frac{2[\frac{1}{2} - \bar{\Phi}(\kappa_u)]}{\kappa_u \phi(\kappa_\ell)}$$

where $\kappa_\ell = y^*$ is the lower-bound solution following (37), κ_u is the upper-bound solution to the optimality equation in (42), and $\kappa_\ell \leq \kappa_u$. Consequently, we have

$$2 \leq \frac{V^u}{V^\ell} \leq \frac{2\phi(0)}{\phi(\kappa_\ell)}$$

and

$$\frac{V^\ell}{V^u} \leq \frac{V^\ell}{V^*} \leq \frac{V^u}{V^*} \leq \frac{V^u}{V^\ell}.$$

Based on the foregoing analysis, we propose to use $\alpha = \frac{1}{2}$ and $\kappa = \kappa_u$, the solution to the upper-bound problem V^u (for $\kappa > 0$), for our approximate solution to the original optimization problem in (34). When $\kappa < 0$, we can still use the solution to V^u as a heuristic, simply because the upper- or lower-bounds in this case will not lead to meaningful solutions. (For instance, the upper-bound problem is $\min_{\kappa \leq 0}(c - p)\kappa$, and thus the solution is: $\kappa^* = 0$ if $c \leq p$, and $\kappa^* = -\infty$ if $c > p$.)

All we need is to remove the constraint $\kappa \geq 0$ from the minimization problem in (41), and then the optimality equation in (42) still applies (with the right-hand side *increasing* in κ for $\kappa \leq 0$). However, it is no longer the case that V^u provides an upper bound. In fact, when $\kappa < 0$, V^u is a *lower* bound of V^*: the constructed linear lower bound to the concave function $\sqrt{n}$, along with a negative κ, implies subtracting *less* from the production costs, which in turn must result in lower lost-sales penalties. Thus, we have $V^u/V^* \leq 1$, and the performance of V^u, as an approximation to V^*, should be comparable to, if not better than, its performance in the case of $\kappa > 0$.

4.2. Incorporating Inventory Cost

Now, we turn to incorporate inventory costs in our analysis. Referring back to (35), let h/N be the holding cost per unit of inventory that is held for one period, with $h > 0$ a constant parameter. The reason for including N in the denominator can be explained as follows. The application context we have in mind is a fast-moving production-inventory system, where the planning horizon is of a fixed length, say one unit, and h is the cost to hold one unit of inventory over this entire horizon. By dividing the horizon into N segments (periods), we are effectively using N as a scaling parameter on the demand. Hence, as previously noted, instead of demand D over the entire planning horizon, we have been considering D_n for each period $n = 1, \ldots, N$, with $D_n \overset{\mathrm{d}}{=} D$ and independent of n. Another way to motivate this approach is due to the fact that, if h is not scaled by N, then the inventory holding cost, because of its cumulative nature over time, will dominate all other costs when N grows.

Therefore, using $\alpha = \frac{1}{2}$, the inventory cost is given by

$$\frac{h}{N}\sum_{n=1}^{N} \sigma\sqrt{n}[G(\kappa) + \kappa] \approx \frac{2}{3}h\sigma\sqrt{N}[G(\kappa) + \kappa]$$

for the lower bound; and similarly, for the upper bound, we have

$$\frac{h}{N}\sum_{n=1}^{N} \sigma\sqrt{n}\left(\frac{1}{\kappa}\left[\frac{1}{2} - \bar{\Phi}(\kappa)\right] + G(\kappa) + \kappa\right) \approx \frac{2}{3}h\sigma\sqrt{N}\left(\frac{1}{\kappa}\left[\frac{1}{2} - \bar{\Phi}(\kappa)\right] + G(\kappa) + \kappa\right);$$

refer to (23) and (25). Upon replacing $\frac{2}{3}h$ by h, the inventory costs derived above are easily incorporated into both the Brownian approximation and the lower- and upper-bound surrogate problems of the previous subsections: simply replace c by $c + h$, and p by $p + h$.

4.3. Asymptotic Equivalence of Lost-Sales and Backorder Models

Next, we consider another type of asymptotic results. Suppose the penalty p is large, say $p \to \infty$, whereas N is fixed. We first ask the question of whether it is possible to keep the expected penalty cost $p\mathsf{E}[L_N]$ finite (for a given N).

To this end, let us consider the lower-bound solution with $\alpha = \frac{1}{2}$ and $\kappa = \kappa_\ell = y^*$; for simplicity, we shall write y instead of y^*. Recall from (37) that $y = \bar{\Phi}^{-1}(\frac{c+h}{p+h})$ when inventory cost is included into the model, and thus $p \to \infty$ is equivalent to $y \to \infty$. Then, for any n, we have

$$p\mathsf{E}[S_n^+] = \sigma\sqrt{n}pG(y) = \sigma\sqrt{n}\frac{cG(y)}{\bar{\Phi}(y)},$$

for which applying l'Hôpital's rule yields

$$\lim_{y\to\infty}\frac{G(y)}{\bar{\Phi}(y)} = \lim_{y\to\infty}\frac{\bar{\Phi}(y)}{\phi(y)} = \lim_{y\to\infty}\frac{1}{y} \to 0.$$

Namely, $p\mathsf{E}[S_n^+] \to 0$ for every n.

In fact, with $S_n^+(y) := \sigma\sqrt{n}(Z-y)^+$, this strategy also leads to $\mathsf{E}[S_n^+(y)] = \sigma\sqrt{n}G(y) \to 0$ as $y \to \infty$ for every n. Because $S_n^+(y)$ is decreasing in y, monotone convergence implies that $S_n^+(y) \to 0$ as $y \to \infty$, for every $n = 1, \ldots, N$. Hence, we must have $L_N(y) = \max_{1\le n\le N} S_n^+(y) \to 0$ as $y \to \infty$, and thus $\mathsf{E}[L_N(y)] \to 0$ as $y \to \infty$. Since the optimal strategy (to the original problem with inventory cost included) can do no worse, we must have

$$\min_{\kappa}\left(\sigma\sqrt{N}(c+h)\kappa + (p+h)\mathsf{E}[L_N(\kappa)]\right)$$
$$\le \sigma\sqrt{N}(c+h)y + (p+h)\mathsf{E}[L_N(y)] = \sigma\sqrt{N}(c+h)y + o(y),$$

where $y = \bar{\Phi}^{-1}(\frac{c+h}{p+h})$ as specified above.

On the other hand, based on the results established earlier, we know that the right-hand side above is also a lower bound of the objective function, because (with κ^* denoting the minimizer of the first expression)

$$\min_{\kappa}\left(\sigma\sqrt{N}(c+h)\kappa + (p+h)\mathsf{E}[L_N(\kappa)]\right)$$
$$\ge \sigma\sqrt{N}(c+h)\kappa^* + (p+h)\mathsf{E}[S_N^+(\kappa^*)]$$
$$\ge \sigma\sqrt{N}(c+h)y + (p+h)\mathsf{E}[S_N^+(y)] = \sigma\sqrt{N}(c+h)y + o(y),$$

where the second inequality follows from y being the minimizer of the lower-bound problem. We therefore have

$$\min_{\kappa}\left(\sigma\sqrt{N}(c+h)\kappa + (p+h)\mathsf{E}[L_N(\kappa)]\right) = \sigma\sqrt{N}(c+h)y + o(y). \tag{46}$$

Next, let us consider the backorder model. Recall that $B_n = S_n^+$ is the number of backlogged units of demand and that $H_n' = S_n^-$ is the inventory, both at the end of period n; refer to Proposition 1. Similar to the lost-sales model in the previous subsection, let $\frac{b}{N}$ and $\frac{h'}{N}$ be the per-unit backlog penalty and inventory holding costs in each period, respectively. Then, the (original) optimization problem can be expressed as

$$\min_{\kappa}\left(c\kappa + \frac{b}{N}\sum_{n=1}^{N}\mathsf{E}[S_n^+(\kappa)] + \frac{h'}{N}\sum_{n=1}^{N}\mathsf{E}[S_n^-(\kappa)]\right) = \sigma\sqrt{N}\min_{\kappa}[(c+h')\kappa + (b+h')G(\kappa)], \tag{47}$$

where $\frac{2}{3}b$ and $\frac{2}{3}h'$ are replaced by b and h' (analogous to the approach taken with h in the previous subsection), respectively, and taking into account $S_n^- = S_n^+ - S_n = S_n^+ + \sigma\sqrt{n}\kappa$. Let y' denote the optimal solution to (47), clearly rendering $y' = \bar{\Phi}^{-1}(\frac{c+h'}{b+h'})$. Suppose we set $b = p$ and $h' = h$, and let $p \to \infty$. Then, $y' = y$ and the left-hand side of (47) has the following asymptotics for large y:

$$\sigma\sqrt{N}\min_\kappa \left(c\kappa + b\mathsf{E}[S_n^+(\kappa)] + h'[S_n^-(\kappa)]\right) = (c+h)y' + o(y') = (c+h)y + o(y). \tag{48}$$

Upon comparing (46) and (48), we have established the following result.

PROPOSITION 11. As $p \to \infty$, the objective values of the lost-sales and backorder models become equivalent in that their ratio approaches 1, provided the lost-sales and backlog penalty costs are both equal to p, and the purchasing and holding costs are also respectively equal in the two models.

5. Conclusions

A general class of high-volume, fast-moving production-inventory problems are studied. Although we focus on the case in which the production is set to be the average demand supplemented by a certain amount of safety stock, our assumptions on this safety stock is quite general. In the asymptotic regime, we identify the right order of the safety stock by studying random walks with power drifts. We also derive bounds and approximations for functionals of key performance measures so that the production planning can be optimized under different settings. Finally, our analysis provides another means and settings for establishing the asymptotic equivalence of lost sale and backorder models, as observed by several authors using different means and settings.

References

[1] ABRAMOWITZ, M. AND STEGUN, I. *Handbook of Mathematical Functions.* Dover, New York, 1965.

[2] ALILI, L., PATIE, P., AND PEDERSEN, J. Representations of First Hitting Time Density of an Ornstein-Uhlenbeck Process, *Stochastic Models*, **21**, 2005, (967-980).

[3] BERTSIMAS, D., AND POPESCU, I. Optimal Inequalities in Probability Theory: A Convex Optimization Approach, *SIAM J. Optim.* vol. 15, No. 3, 780-804, 2005.

[4] BREIMAN, L. First Exit Time from a Square Root Boundary. *Proc. 5th Berkeley Symp. Math. Stat. Prob.,* **2**, 9-16, 1966.

[5] CERBONE, G., RICCIARDI, L.M. AND SACREDOTE, L., Mean Variance and Skewness of the First-Passage Time for the Ornstein-Uhlenbeck Process, *Cybernetics and Systems*, **12**, 395-429, 1981.

[6] CHEN, H., AND YAO, D.D., *Fundamentals of Queueing Networks*, Springer, New York, 2001.

[7] CHOU, M., CHUA, G., TEO, C., AND ZHENG, H., Design for Process Flexibility: Efficiency of Long Chain and Sparse Structure, *Operations Research*, **58**(1), 43-58, 2011.

[8] GOLDBERG, D.A., KATZ-ROGOZHNIKOV, D.A., LU, Y., SHARMA, M. AND SQUILLANTE, M.S., Asymptotic Optimality of Constant-Order Policies for Lost Sales Inventory Models with Large Lead Times, *Preprint*, 2012.

[9] GÖING-JAESCHKE, A. AND YOR, M., A Clarification Note About Hitting Times Densities for Ornstein-Uhlenbeck Processes, *Finance Stochastics*, **7**, 413-415, 2003.

[10] HUH, T., JANAKIRAMAN, G., MUCKSTADT, J. AND RUSMEVICHIENTONG, P., Asymptotic Optimality of Order-Up-To Policies in Lost Sales Inventory Systems, *Management Sciences*, **55**(3), 404-420, 2009.

[11] HUH, T., AND JANAKIRAMAN G., On the Optimal Policy Structure in Serial Inventory Systems with Lost Sales, *Operations Research*, **58**(2), 486-491, 2010.

[12] KARLIN, S., SCARF, H., Inventory models of the Arrow-Harris-Marschak type with time lag, *Studies in the Mathematical Theory of Inventory and Production,* Stanford University Press, Palo Alto, CA, 155–178, 1958.

[13] LEBLANC, B., RENAULT, O., AND SCAILLET, O., A Correction Note on the First Passage Time of An Ornstein-Uhlenbeck Process to A Boundary, *Finance Stochastics*, **4**, 109-111, 2000.

[14] LEVI, R., JANAKIRAMAN, G. AND NAGARAJAN, M., A 2-Approximation Algorithm for Stochastic Inventory Control Models with Lost-Sales, *Mathematics of Operations Research*, **33**(2), 351-374, 2008.

[15] MORTON, T. E., The near-myopic nature of the lagged-proportional-cost inventory problem with lost sales. *Oper. Res.,* **19**(7) 1708–1716, 1971.

[16] NOBILE, A.G., RICCIARDI, L. M. AND SACERDOTE, L., Exponential Trends of Ornstein-Uhlenbeck First-Passage-Time Densities, *Journal of Applied Probability,* **22**, 360-369, 1985.

[17] RICCIARDI, L., SATO, S., First-Passage-Time Density and Moments of the Ornstein-Uhlenbeck Process, *Journal of Applied Probability*, **25**(1), 43-57, 1988.

[18] ROSS, S.M., *Stochastic Processes,* Wiley, New York, 1996 (2nd ed).

[19] SLAMINEN, P. On the First Hitting Time and the Last Exit Time for a Brownian Motion to/from a Moving Boundary, *Adv. Appl. Prob.* **20**, 411-426, 1988.

[20] SPITZER, F., A Combinatorial Lemma and Its Application to Probability. *Transactions AMS,* **82** (1956), 323-339.

[21] SATO, S., On the Moments of Firing Interval of the Diffusion Approximated Model Neuron, *Math. Biosci.*, **39**, 53-70, 1978.

[22] SIEGERT, A.J.F. On the First Passage Probability Problem. *Physice Review*, **81**, 617-623, 1951.

[23] ZIPKIN, P. On the Structure of Lost-Sales Inventory Models, *Operations Research*, **58**, 2008.